Forecasting and Verifying in a Field Research Project: DOPLIGHT '87

CHARLES A. DOSWELL III

NOAA/National Severe Storms Laboratory, Norman, Oklahoma 73069

JOHN A. FLUECK

University of Colorado/CIRES and NOAA/ERL/Environmental Sciences Group, Boulder, Colorado 80303

(Manuscript received 19 April 1988, in final form 17 January 1989)

ABSTRACT

Verification of forecasts during research field experiments is discussed and exemplified using the DOPLIGHT '87 experiment. We stress the importance of forecast verification if forecasting is to be a serious component of the research. A direct comparison and contrast is done between forecasting for field research and forecasting in the operational sense, highlighting the differences between them. The verification of field research program forecasting is also different from that done in operations, as a result of those forecasting differences.

DOPLIGHT '87 was a field project conducted jointly by the National Severe Storms Laboratory and the Oklahoma City National Weather Service Forecast Office, and is described in detail. During the experimental design, special attention was given to forecast design, to ensure that verification would be unambiguous and that the data collected would be appropriate for validating the forecasts. This a priori design of the forecasts to consider proper objective verification is, we believe, unique among research field programs. The forecast evaluation focuses on the contingency table and summary statistics derived from it, as treated in a companion paper by Flueck (1989; hereafter referred to as Flu89).

Results are interpreted in terms of their implications for future field research experiments and for operational forecasting. For example, it is noted that DOPLIGHT '87 forecasts of convective potential were nearly constant from the evening before an anticipated operational day to about local noon on that day. This suggests that convective storm field research operational decisions could be made as early as the evening before an anticipated operational day with negligible loss of skill. Summary measures of the forecast verification suggest that the DOPLIGHT '87 forecasters demonstrated skill roughly comparable to the forecasters at the National Severe Storms Forecast Center in issuing outlooks of convective potential. The requirement for time to assimilate the most recent data is noted both for field experiments and for operations, and some discussion of the potential impact of new data acquisition and processing systems is offered.

1. Introduction

While most field programs use forecasting in support of their operations, verification is rarely performed on those forecasts.[1] Therefore, many field program forecasting efforts have not been designed with verification requirements in mind (see Doswell et al. 1986). Although forecasts may be required for successful operation of the field program, the forecasting itself is often not considered to be an important scientific element in the project. If one is serious about forecasting, one also should be serious about verification of those forecasts (e.g., Flueck 1987), since there always is room for improvement. It is through verification that one learns what is going wrong and, just as importantly, what is going right with the forecasting effort. In our view, this should be the primary purpose for verification.

During the spring of 1987, the National Severe Storms Laboratory (NSSL) and the Oklahoma City National Weather Service Forecast Office (located in Norman, Oklahoma) collaborated on a field program, called DOPLIGHT '87.[2] This joint effort was designed to serve a broad range of scientific and operational objectives (see Forsyth et al. 1988), one of which was to assess skill at forecasting mesoscale convective weather systems. An effort was made to include elements in the DOPLIGHT '87 forecasts that could be subjected to objective evaluation. A discussion of the unique as-

[1] Rare exceptions to this include the Canadian Atlantic Storms Program (MacDonald et al. 1988) and the Sierra Cooperative Pilot Project (Flueck and Reynolds 1986).

Corresponding author address: Charles A. Doswell III, NOAA/National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.

[2] DOPLIGHT '87 is a continuation of a series of experiments during which Doppler radar and lightning ground strike information were brought into an operational forecast office. Hence, the name DOPLIGHT is derived from DOPpler radar and LIGHTning data.

pects of forecasting for field research programs is given in section 2 of this paper, including specific aspects of DOPLIGHT '87 forecasting experimental design. Section 3 details aspects of verification for such forecasting experiments, once again giving specific attention to the verification of DOPLIGHT '87 forecasts. The results of that verification are presented in section 4, including some implications for operations, and conclusions are drawn in section 5.

2. Characteristics of forecasting for field research

a. Basic issues

It is useful to understand how forecasting for field research studies compares to that done operationally within the National Weather Service. Field research programs typically are space- and time-limited efforts (e.g., within and near the state of Oklahoma from 15 March to 15 June), within which many expensive resources such as Doppler radars, research aircraft, and mesonetworks are concentrated for purposes of collecting specialized research data sets during specific types of weather events such as intense convective storms. In most such experiments at least some of the data collection is done by special sensing systems that may be untried in field conditions (e.g., wind profilers, disdrometers, etc.). Although the weather of concern for field programs may be limited to only a few specific types, as opposed to the full range of weather events treated by operational forecasters, demands for time and space specificity are usually greater for field researchers than for operational forecasters. A great deal of expensive and quite limited resources may be at stake with each forecast, creating considerable pressure on the forecasting team to predict accurately the occurrence of the desired events. Unfortunately, it is common for the personnel in such research projects to be relatively inexperienced at weather forecasting, with the forecasters often being research scientists.

Forecasting in field research (like operational forecasting) may involve both regularly-scheduled products and a host of forecasts (or nowcasts[3]) made as needed. The egregious practice of "second guessing" the forecasts is probably even more common in field research than in operational forecasting. Curiously, this practice is not much mitigated even in programs where the forecasters and researchers exchange roles on a regular basis. While forecasting may be a big factor in determining the success or failure of the field research, the project director is normally the one making final decisions about the program (see Doswell et al. 1986). In operational forecasting, the forecast itself is a decision, whereas in field research, the forecast is simply one of the inputs considered when arriving at a decision. Within the research team, internal conflict and tension arising from disagreements about the forecast can affect the decision-making process adversely; there is a tendency for dominance of personality and prestige within the research community to be more influential than meteorological reasoning.

In some programs, the data from special instrumentation (e.g., special rawinsondes or wind profilers) is fed back to the forecast center in real time, creating additional problems for the forecaster. The experimental observing systems may be difficult to interpret properly, especially if little or no experience with them is available (e.g., vertical velocities derived from steerable Doppler radar wind profiles in clear air). Although the special data may be of a sort familiar to the forecasters (as in special rawinsondes), the data may be unrepresentative. Often, the new data systems produce a great deal of data in a short time (e.g., meteorological satellites or Doppler radars), potentially creating "information overload" in which the flood of new information cannot be assimilated by the forecaster (see also MacDonald et al. 1988). Beyond the possible flood of raw data from the new remote sensing systems, a plethora of derived products from those same raw data can add to the problem. Many field programs involve mobile field observing teams whose requirements for weather information and coordination creates additional demands on the forecast team. (In some programs, the coordination and nowcasting duties are accomplished by separate groups; in other programs, those duties are shared.) The feedback about ongoing events from field teams can be a valuable input to the forecaster, but their need for nowcasts and external direction can conflict with the forecaster's responsibility to forecast.

b. Relation between forecasters and forecast users

As already noted, the forecast in a field research experiment is only one of a number of inputs to the decision maker(s) directing the project. Moreover, the users of the forecasts usually receive the forecast directly from the forecaster in face-to-face briefings. (This certainly was the case for DOPLIGHT '87.) Many of the forecast users may be as qualified to forecast as the person designated as forecaster, with users and forecasters exchanging roles in some projects. Even those users not qualified to forecast (engineers, technicians, pilots, etc.) often have technical skills comparable to those of the forecaster. This can result in a different relationship between forecaster and user than that which is characteristic in operational forecasting. While the user of operational forecasts may wish to know generally how well the forecasts verify for decision making, the performance of field program forecasts is

[3] For purposes of this paper, a nowcast is defined to be a short-range (e.g., 0–2 h) forecast concerning a weather event already in progress, usually based on linear extrapolation. If the event of interest is not yet in progress, or if the extrapolation involves nonlinear behavior, then it cannot be a nowcast.

usually *directly* related to the success or failure of the intended research. Hence the user of such forecasts has a large vested interest in their success far beyond the typical user of operational forecasts.

Naturally, the actual weather events have a considerable impact on the outcome of a field experiment (as noted in Doswell et al. 1986). Given that events of interest to researchers often are perceived by the public as "bad" weather, the successful outcome of a field experiment usually depends on the occurrence of "bad" weather within the limited temporal and spatial confines of the project. Since forecasting "good" weather is generally uninteresting, the interest of the forecasters and the users of the forecast usually coincide in wanting "bad" weather. This coincidence of interests between forecasters and the forecast users in research field projects is somewhat in contrast to the situation in operations.

TABLE 1. Forecast products, issue times, and valid times. All times are given in local station time (L), which is in the central time zone; this is standard time, switching to daylight time on 5 April.

Product	Issue time	Valid time
1. Advance outlook	1700–2359L	0900–2300 L
*Categorical go/no-go		(next day)
2. Morning update	0900 L	0900–2300 L
*Categorical go/no-go		
3. Noon outlook	1200 L	1200–2300 L
a. Categorical go/no-go		
b. Mesocyclone yes/no		
c. Go probability		
d. Mesocyclone probability		
e. OTO narrative discussion		
4. Afternoon update	1500 L	1500–0600 L
a. Convective mode		
b. OTO narrative update		

c. *Design of the DOPLIGHT '87 experiment*

The primary focus of the DOPLIGHT '87 experiment was on the use of Doppler radars in an operational setting. It also was a chance to evaluate forecasts of convective weather quantitatively. Thus, a substantial effort went into designing forecast products to meet the requirements of this study; viz., the forecast products had to include precise statements that could be validated using information that would be generated by the DOPLIGHT '87 project. The rationale for each forecast product is given henceforth. There also were products associated with DOPLIGHT '87 operational needs that were included in the forecast routine, but which could not be verified in rigorous fashion (e.g., narrative weather discussions) and, hence, will not be discussed in this paper.

During each day of the experiment, which ran from 15 March to 15 June 1987, it was expected that the forecast team on duty (consisting of a lead forecaster and an assistant) would fill out a forecast sheet in addition to any duties required by the experimental operation.[4] Forecast products are listed in Table 1. Forecast duty hours were left up to individual forecasters, subject to the requirements of forecast issuance times and the need to support DOPLIGHT '87 operations during convective weather.

For products 1, 2, and 3, the verification area included all Oklahoma and North Texas counties within 230 km of the Norman, Oklahoma Doppler radar, plus any counties in the Oklahoma City National Weather Service Forecast Office (NWSFO) area of warning responsibility. These are shown in Fig. 1. Product 4 was verified over the entire state of Oklahoma, including the Panhandle. The Oklahoma Thunderstorm Outlook (OTO) is an operationally produced narrative discussion issued by the Oklahoma City NWSFO for public dissemination. During DOPLIGHT '87, the OTO was prepared jointly by the NWSFO and the DOPLIGHT forecast team. This collaboration also included a daily conference call with the Severe Local Storms (SELS) Unit of the National Severe Storms Forecast Center (NSSFC), Kansas City, Missouri; the Fort Worth, Texas NWSFO forecaster responsible for preparing a product comparable to the OTO for North Texas; and the severe weather forecaster at U.S. Air Force Global Weather Central (Offutt Air Force Base, Nebraska).

The convective mode forecast was an effort to determine how well the forecasters could anticipate the *dominant* mode of convection during the late afternoon and night. The forecasters were to choose from the list of modes shown in Table 2. The *observed* convective mode used in verifying these mode forecasts was determined by the following procedure. Long after the completion of the experiment, two NSSL meteorologists examined the relevant hard copy satellite imagery for every operational day and each made an independent determination of the observed mode, without knowledge of the forecasts. Any discrepancies between their independent determinations were then reconciled—again without knowledge of the forecasts—to arrive at the final mode determination for each day of the experiment.

At the end of each operational day, the DOPLIGHT Doppler Radar Interpreter made a decision about whether or not field intercept teams *should* have been deployed, regardless of the actual deployment decision. It was the mission of the intercept teams to document any severe weather phenomena within the operations area shown in Fig. 1. Thus, the Doppler radar interpreter based his assessment of the need for intercept operations on whether or not "suspicious" radar signatures (clearly, a subjective assessment) were present

[4] The forecasters gave a daily briefing at noon for the DOPLIGHT '87 operational staff and, in the event of field operations, were to provide field teams with weather information and to coordinate field team activities with the rest of the DOPLIGHT '87 operation.

TABLE 2. Convective mode types offered to forecasters on their forecast sheets. The definition of a mesoscale convective complex (MCC) follows the criteria set in Maddox (1980).

1. No deep, moist convection
2. One or more isolated convective storms
3. One or more storm complexes (other than "5")
4. One or more squall lines
5. One or more MCC's (Maddox 1980 criteria)
6. Mode ___ changing to mode ___ (specify)

during the operational day for which chase team documentation would have been useful. Note that such a predictand differs significantly from one based on reports of severe weather occurrence or nonoccurrence.

It is worth taking some time to explain the reasons for using this approach instead of simply using the severe weather occurrence data and forecasting the occurrence of severe weather events. Although the interpretation of the radar is subjective, the radar provides essentially uniform data everywhere within the forecast area which, indeed, was chosen specifically to match the radar data coverage. Severe-weather occurrence data require someone to be present in the vicinity of the event, to interpret the event correctly, and to report it. Thus, severe weather occurrence data are notoriously unreliable (see Doswell and Burgess 1988; Kelly et al. 1985), even when chase teams are employed. We chose this scheme to avoid the pitfalls of depending on severe weather occurrence data, in spite of the subjective nature of the radar interpretation to our concern to develop verifiable forecast products in advance of the project; the experiment was designed to give us an unambiguous determination of "go" or "no-go" decisions. Hence, one should not interpret these data as equivalent to the occurrence or nonoccurrence of severe weather. Also, in the same way, the Doppler radar interpreter determined whether or not a mesocyclone was detected on an operational day within 230 km of Norman. Given the definition of a mesocyclone as a certain type of Doppler radar signature, this too was as unambiguous as possible. This information was used to verify the mesocyclone forecasts.

We should note briefly the make-up of the experimental forecast team. Members were drawn exclusively from the NSSL staff (see Forsyth et al. 1988 for a listing of the team members). Their experience at forecasting varied rather widely, with the most experienced generally serving as lead forecaster and the least experienced serving as assistants. Except for some familiarization training with NWSFO equipment, there was no special forecast training given to the team.

3. Characteristics of verification for research studies

a. Basic issues

Verification of operational forecasts is also different from verification of research project forecasts. Operational forecasting verification typically is confined to point forecasts of, for example, temperature and precipitation, whereas the forecast problem in field programs typically is whether or not an event will occur within the operational area. Thus, verification of the forecasts in a field experiment is most logically done for areas rather than points. The validating data are usually the same datasets that are collected for research purposes, perhaps supplemented by operational datasets. In contrast to operational verification involving

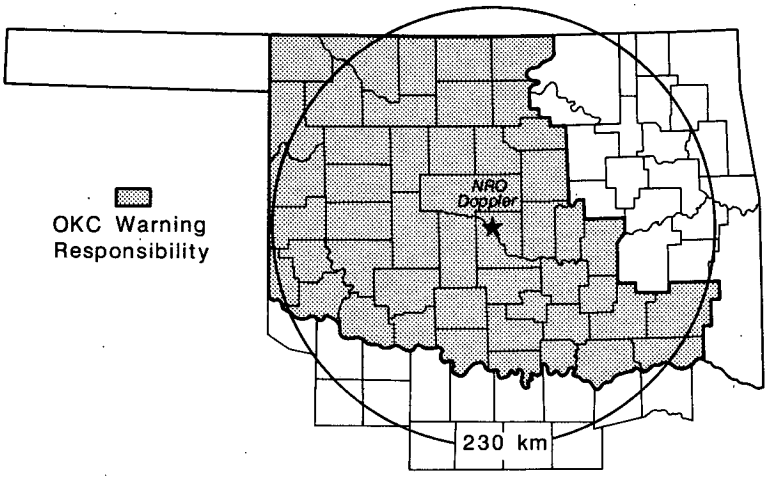

FIG. 1. Map showing the DOPLIGHT '87 forecast and verification area. All counties in Oklahoma and north Texas touched by the circle within 230 km of the Norman (NRO) Doppler radar are outlined with fine lines. Stippling inside the heavy line denotes the counties for which the Oklahoma City National Weather Service Forecast Office (OKC) has warning responsibility. The forecast area for the DOPLIGHT '87 forecasts is the union of these two regions.

relatively sparse data over a large area, the research forecasts have relatively dense data over a limited area. While the research data may have higher resolution than operational data, the experimental nature of the observing systems may yield information of less known quality.

Furthermore, issuance and verification of operational forecasts is more or less a standardized procedure. National and regional procedures are mandated at levels well beyond that of the forecaster. In contrast, verification of forecasts issued during a field program can be much more flexible than operational verification. When a methodology for verification is developed, the method itself makes a statement about what constitutes a correct forecast. Thus, the verification methodology is of considerable significance.

b. Verification methodology

There are many forecast verification schemes available in the literature (see reviews by Brier and Allen 1951; Panofsky and Brier 1968; Dobryshman 1972) from which one might choose. We believe that the principal purpose of verification is evaluation, i.e., insight into what is right and what is wrong about the forecasts (see the excellent discussion in Panofsky and Brier 1968, p. 192 ff.), rather than mere production of verification statistics for ranking of relative performance. Therefore, our approach to verification has concentrated on a few simple methods, as presented by Flu89, designed to give the DOPLIGHT '87 participants some insight on their forecasting performance.

From one point of view, there are two types of forecast products: categorical and probabilistic. Typically, the statistical techniques for evaluating these two forecast types are different. As indicated in Flu89, however, this distinction is illusory. Categorical forecasts are simply dichotomous probabilities, viz, zero and unity. Conversely, forecasts with more than two probability categories are called *polychotomous* probabilities (see Flu89) which can be combined with thresholding techniques to convert them to dichotomous probabilities if desired.

The DOPLIGHT '87 forecasting experiment included a predominance of dichotomous predictions, but certain forecasts also employed concurrent polychotomous probability estimates. Our primary approach to verification of the dichotomous forecasts is through the traditional contingency table and its associated graphical and algebraic summary measures (see Brownlee 1965; Wilson and Flueck 1986). These include the probability of detection (POD), the false alarm ratio (FAR) and the critical success index (CSI) as defined by Donaldson et al. (1975); the probability of false detection (POFD) and the true skill statistic (TSS) as presented in Flu89; and concepts of signal detection theory (e.g., Mason 1982).

We also will present some conventional verification measures, for reference to past verification in operational forecasting, including the Brier score, skill score, and bias. The formulas for these are presented in appendix A.

4. DOPLIGHT '87 results

a. Quantitative evaluation

The 1987 spring severe weather season in Oklahoma was characterized by a dearth of severe weather (Ostby et al. 1988). For the first time since reasonably reliable tornado reporting began [i.e., since 1950 (see Kelly et al. 1978)], the entire month of April passed without a single reported tornado in the state. During all 93 days of the experiment, there was not a single mesoscale convective complex meeting the Maddox (1980) criteria within the DOPLIGHT '87 experimental area. Therefore, a large percentage of the forecasts were for no severe weather. While this seems anomalous in view of the historical record, 1988 already has shown a similar lack of "normal" severe weather in Oklahoma.

Products 1, 2, and 3 (i.e., the advance outlook, morning update, and noon outlook) all included cat-

TABLE 3. Contingency table and summary statistics for Advance Outlook.

	Observed				
Predicted	Go	No-go	Total	Statistics	
Go	18	4	22	POD = .69	POFD = .06
No-go	8	61	69	FAR = .18	
Total	26	65	91	CSI = .60	
				TSS = .63	

TABLE 4. Contingency table and summary statistics for Morning Update.

	Observed				
Predicted	Go	No-go	Total	Statistics	
Go	17	7	24	POD = .65	POFD = .10
No-go	9	58	67	FAR = .29	
Total	26	65	91	CSI = .52	
				TSS = .55	

TABLE 5. Contigency table and summary statistics for Noon Outlook's categorical go/no-go forecast.

	Observed				
Predicted	Go	No-go	Total	Statistics	
Go	19	5	24	POD = .73	POFD = .08
No-go	7	60	67	FAR = .21	
Total	26	65	91	CSI = .61	
				TSS = .65	

egorical forecasts of "go" conditions, as defined in section 2c. Although there are 93 days in the experimental period, the forecasts for two obviously nonconvective days early in experiment were not recorded as a result of a misunderstanding about procedures. The contingency tables for these three separate forecasts are given in Tables 3, 4, and 5; note that products 1 and 3 verified rather similarly, in spite of their considerable difference in issue time.

While this experiment was exploratory [as described in Flueck (1986)], we have calculated the large sample standard deviation (see Dixon et al. 1985) of the TSS scores for subjective guidance on "signal-to-noise" interpretation. As noted in Flu89, the TSS is a measure of the association between the predictions and the observations. The large sample standard deviation (or standard error) statistic estimates the variability associated with noise: the smaller its value, the lower the noise level. Some assumptions (e.g., a large sample size, normality of the underlying distribution, etc.) are required if one is to use the large sample standard deviation for estimating the variance of the TSS (see Flueck 1987). There is no way of determining from the DOPLIGHT '87 dataset how valid these assumptions are; however, the large sample standard deviation values are 0.10, 0.10, and 0.09, for the advance outlook, morning update, and noon outlook, respectively. On the basis of estimated signal-to-noise ratios, therefore, it appears that the TSS values are substantial and the Morning Update forecast has the lowest summary value (i.e., 0.55).

For comparison purposes, we have used the convective outlook (AC) products issued by the NSSFC (see Weiss 1977) to make a second set of "go" forecasts for each day of the experiment. This was done by determining whether or not the area enclosed by the graphic outline of the NSSFC Outlook included any nontrivial fraction of our forecast area—if so, that was considered a "go" forecast by SELS.[5] There are three such outlooks issued daily: an early AC (issued at 0700 UTC), a morning AC (issued at 1500 UTC), and a noon AC (issued at 1900 UTC). In addition to these AC products valid on the day of issue (day-1), there are two second-day outlooks [issued at 0800 UTC (the day-2 early AC), and at 1800 UTC (the day-2 noon AC)] valid for the day after issuance.

The verification statistics for the ACs are given in Table 6. Both sets of ACs show improvement during the course of a forecast day. However, note that the day-1 early AC product actually shows a small decrease in skill, as measured by the TSS, over the day-2 noon AC forecast. By comparison, our advance outlook product is superior to all but the noon AC on day-1 (Table 7, the only AC for which the contingency table is shown), but our morning update is inferior to the SELS product issued at a comparable time. Note that our forecasts only covered a limited region in contrast to the nationwide area of responsibility assumed by NSSFC and the DOPLIGHT '87 forecast team did not have the wide range of additional duties usually borne by forecasters in a local weather office (see Doswell 1986), apart from nowcasting for (and directing) the chase teams.

The polychotomous probability forecasts issued as part of the noon outlook have a Brier score (described in appendix A) of $B = 0.12$; this suggests that the probability forecasts did rather well. For comparison purposes, typical Brier scores for National Weather Service precipitation probability forecasts range from roughly 0.05 to 0.15 (M. Foster, personal communication). The forecast probabilities averaged 25.4%, compared to an observed "go" day frequency of 28.6%. The bias, $b = -11.2\%$, indicates a slight tendency for underforecasting.

Using the observed "go" day frequency as a constant "climatology" forecast yields an average Brier score of 0.20, giving a forecast skill score (which can be interpreted as a percentage improvement over climatology) of about 40%. Again for comparison purposes, National Weather Service precipitation probability forecasts show a typical range of 25%–50% improvement over climatology, depending on the season. Note that using the observed frequency is not the same as using a long-term climatological frequency in calculating the skill score. Given that a "go" forecast is subtly different from a forecast of severe weather occurrence (see above), we felt that it would be inappropriate to use climatological severe weather occurrence data for this purpose.

We can gain additional insight into the quality of the polychotomous "go" probability forecasts by plotting a so-called reliability diagram (Fig. 2)—see Sand-

TABLE 6. Verification statistics for NSSFC Convective Outlook (AC) products converted to "go" forecasts for DOPLIGHT '87 operations.

	NSSFC product				
	POD	POFD	FAR	CSI	TSS
			Day-1		
Early AC	.69	.17	.38	.49	.52
Morning AC	.73	.14	.32	.54	.59
Noon AC	.77	.12	.29	.59	.65
			Day-2		
Early AC	.62	.13	.33	.47	.49
Noon AC	.65	.10	.29	.52	.55

[5] As noted above, there were 2 days during the experiment on which we failed to issue forecasts. Although ACs were issued on all 93 days of DOPLIGHT '87, the SELS forecasts were not verified for the 2 days on which we failed to document our forecasts. Thus, the statistics in Table 6 are comparable to those calculated for DOPLIGHT '87 forecast products.

TABLE 7. Contingency table and summary statistics for the Noon SELS Day One Outlook product with respect to DOPLIGHT '87 forecast area.

	Observed			
Predicted	Go	No-go	Total	Statistics
Go	20	8	28	POD = .77 POFD = .12
No-go	6	57	63	FAR = .29
Total	26	65	91	CSI = .59
				TSS = .65

ers (1973) or Murphy and Daan (1985, p. 415). In this plot, the observed relative frequency of a hit within each forecast probability category is plotted against forecast probability, with the 45° diagonal line representing perfect reliability. It is evident that the DOPLIGHT '87 "go" probability forecasts did not achieve perfect reliability. However, if the forecasts from 10% to 49% are aggregated into one broad category, there were 17 cases with an average probability of 23% versus an observed "go" frequency of 24%. Similarly, the 25 forecasts in the 50%–100% range had an average probability virtually equal to the observed frequency of 80%. Thus, in this broader sense, the forecasts ended up being quite reliable. Only 4% of the zero probability forecasts were in error, which also turns out to be typical of zero precipitation probability forecasts, (Sanders, personal communication).

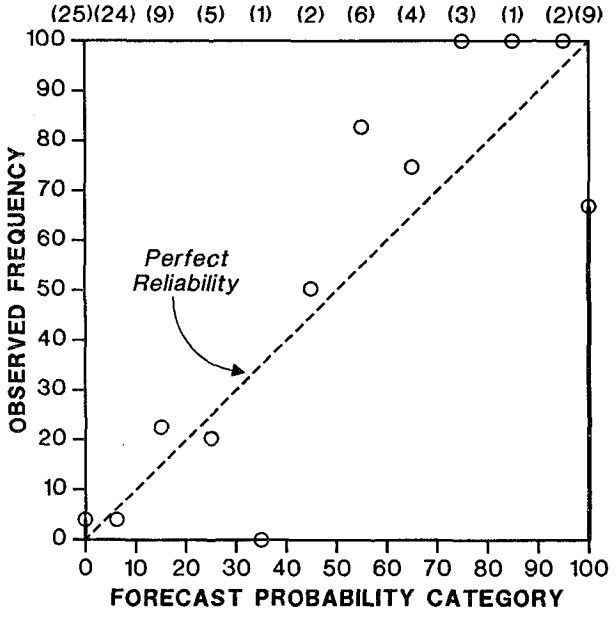

FIG. 2. Reliability diagram for the noon Oklahoma thunderstorm outlook (OTO) "go" probability forecasts. The categories are as described in Table 8, and the dashed line indicates the hypothetical perfect reliability line, along which the observed frequency equals the forecast probability. The numbers in parentheses above each category indicate the number of forecasts made within that category during the experiment.

TABLE 8. Results of testing various thresholds for converting polychotomous "go" probabilities of the Noon Outlook to dichotomous "go" forecasts, forming the basis for the plot in Fig. 3. By "number" is meant the number of forecasts falling within a probability category, while a "hit" is defined as a "go" day occurring within that category. See text for discussion.

Probability category	Number	Hits	POD	POFD	TSS
1.00	9	6	.23	.04	.19
.99–.90	2	2	.30	.04	.26
.89–.80	1	1	.34	.04	.30
.79–.70	3	3	.46	.04	.42
.69–.60	4	3	.57	.06	.51
.59–.50	6	5	.76	.07	.69
.49–.40	2	1	.80	.09	.71
.39–.30	1	0	.80	.10	.70
.29–.20	5	1	.84	.16	.68
.19–.10	9	2	.92	.27	.65
.09–.01	24	1	.96	.63	.33
.00	25	1	1.00	1.00	.00
Totals	91	26			

The polychotomous probability forecasts can be converted to dichotomous forecasts by using thresholds, and then plotting the results on an relative operating characteristic (ROC) diagram (see Mason 1982, or Flu89); i.e., if one selects a threshold probability, say P, then all probability categories meeting or exceeding P are considered "go" forecasts, while all others are "no-go" predictions. For each threshold, the POD and POFD and, hence, the TSS can be found. The results of this process are shown in Table 8 and the associated ROC plot is presented in Fig. 3. For example, in Table 8, for the threshold category of 0.79–0.70, there were 12 hits (6 + 2 + 1 + 3) on "go" forecasts and, thus, there were 14 (i.e., 26 minus 12) misses, giving a POD of 12/26 = 0.46. Correspondingly, for that threshold category, there were 65 "no-go" days (i.e., 91 minus 26) and three (i.e., 15 minus 12) "go" forecasts that failed to verify, giving a POFD of 3/65 = 0.04.

While the reader is urged to consult Mason (1982) for details, some explanation of the ROC diagram is useful here. The ROC curve is designed to assist in the use of verification statistics for decision making, an inherently dichotomous process. In general, decision making involves comparing the probability of correctly detecting an event with the probability of falsely predicting it. If the ROC curve for a particular set of forecasts lies along the diagonal (labeled TSS = 0 on Fig. 3), it indicates no forecasting skill, since a predicted event has an equal probability of being false and being correct. Forecasting skill is associated with ROC curves to the upper left of the diagonal. If the ROC curve has an abrupt change of slope, this suggests where one

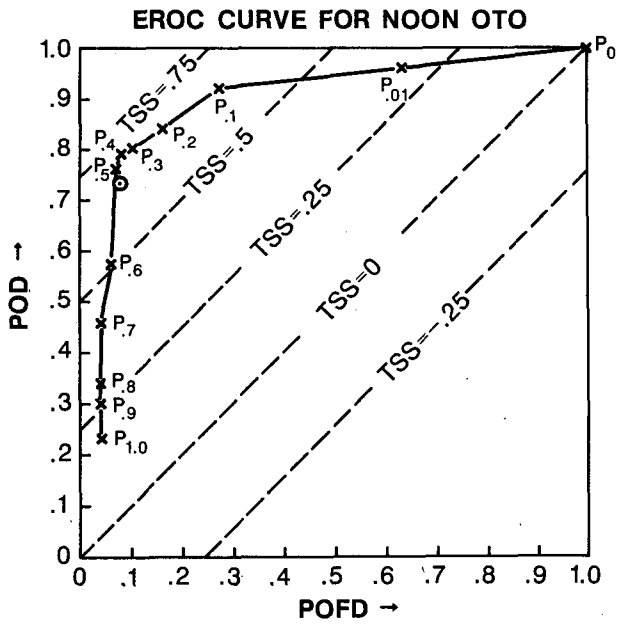

FIG. 3. Empirical relative operating characteristic (EROC) curve (solid line) for the noon OTO forecasts. The x's mark the (POFD, POD) coordinates for dichotomous forecasts using the indicated probability categories (P_i) as the threshold for converting from polychotomous to dichotomous probabilities. Each of the P_i's are labeled with the lowest probability within the category. Also shown by the dashed lines are the TSS contours on this diagram. The circled dot indicates the (POFD, POD) coordinates of the separate dichotomous forecasts also issued with the noon OTO (see text).

would establish thresholds for dichotomous decision making.[6]

The maximum TSS (i.e., 0.71) is achieved within the 0.49–0.40 probability category and, interestingly, the TSS determined from the categorical forecasts issued at the same time (i.e., 0.65) turned out to be fairly close to this maximum value. To this extent, the dichotomous and polychotomous probability forecasts issued by the forecast team are reasonably consistent. The TSS values show a fairly broad maximum from 0.55 to 0.25 probabilities. This suggests that a user of these polychotomous forecasts should be advised to consider it a "go" forecast if the forecast probability is $\geq$0.25. The relative dearth of forecasts in the probability categories from 20% to 90%, however, indicates caution in broadly applying these results.

The paucity of severe weather during the DOPLIGHT '87 experiment is reflected most clearly in the mesocyclone forecasting results shown in Table 9. Thus, during the entire 3 months of DOPLIGHT '87 there were only 6 days on which one or more mesocyclones were documented within the operations area. The summary measures suggest that the dichotomous mesocyclone forecasts showed some skill, but the large sample standard deviation of the TSS is relatively high (i.e., 0.19). Hence, we have decided not to pursue the mesocyclone forecasts beyond presenting these simple results.

As already noted, during the duration of the experiment, there was not a single example of convective mode 5—a mesoscale convective complex meeting the Maddox (1980) criteria. It is of some consolation to the forecasters that this mode also never was forecast. Apart from difficulties encountered in determining the observed mode, some questions also arose about how to score mode 6, which involved a change in the dominant mode. Most convection begins as rather isolated convective elements, which may or may not evolve into other modes. We emphasized to the forecasters that this initial evolution was not to be considered a mode change. However, if the forecast was for a mode change, but one or both of the specified modes were wrong, could this be considered a correct forecast in some sense? Moreover, although exact criteria were given for the MCC mode, equivalently precise specification never was made for the other modes. Although the forecasters were aware that the mode determination was to be via satellite imagery, did radar appearance of the storms influence the *perceived* mode on each convective day? In retrospect, unfortunately, it is clear that the design of this part of the experiment was faulty and better preexperiment training on what was expected would have been helpful.

Nevertheless, we will present the results briefly. Doing so offers us the chance to illustrate the evaluation of a $k \times k$ contingency table (see appendix B), an extension of our 2×2 table evaluations. For the DOPLIGHT '87 convective mode forecasts, the mode-6 verification dilemma described above was resolved in the most generous fashion; i.e., if the observed mode involved a change, this was considered a hit even if one or both of the specified modes were in error. Note that since the mode 5 category was neither forecast nor observed, it has been omitted from the tabulated results; therefore, the table is 5×5.

TABLE 9. Dichotomous mesocyclone forecast contingency table and summary statistics.

	Observed				
Predicted	Yes	No	Total	Statistics	
Yes	4	5	9	POD = .67	POFD = .06
No	2	79	81	FAR = .56	
Total	6	84	90	CSI = .36	
				TSS = .61	

[6] A change in slope of the ROC curve is but one factor in establishing thresholds and it ignores the practical value of forecasting decisions (see Thompson and Brier 1955), which involves the potential costs and benefits associated with establishing thresholds. If one is concerned only with forecast *verification* accuracy, then the ROC curve behavior is a valuable tool in guiding the thresholding process.

TABLE 10. Convective mode forecast contingency table and summary statistics. Note that mode 5, which was neither observed nor forecast, has been excluded from the table.

Forecast mode	Observed mode 1	2	3	4	6	Total
1	30	7	2	1	1	41
2	3	10	0	3	0	16
3	1	6	1	2	3	13
4	0	2	0	2	3	7
6	1	6	1	0	4	12
Total	35	31	4	8	11	89

Statistics

POD_1 = .86	POD_2 = .32	POD_3 = .25	POD_4 = .25
FAR_1 = .27	FAR_2 = .38	FAR_3 = .92	FAR_4 = .71
POD_6 = .36	TSS = .36		
FAR_6 = .67			

The POD and FAR for each forecast category in the table are calculated as described in appendix B. Results of this are shown in Table 10. Note that with the exception of mode 1 (the "no deep, moist convection" mode), the five individual POD and FAR values show little evidence of skill. The generalization of the TSS (described in appendix B) uses the data of Table 10, giving a value of 0.36, which is not particularly encouraging. In spite of the flaws in this part of the experiment, and the overall dearth of severe events, it still seems reasonable to suggest that our forecasters were not notably skillful in forecasting the convective mode, especially considering that mode 6 was given the most generous possible interpretation; i.e., it seems that most of the skill in convective mode forecasting was in distinguishing between days with, at most, isolated convection from those with organized convective storms. The demonstrated ability to predict the observed form of organization was, apparently, quite limited.

b. *Discussion of the results*

The quantitative verification of forecasts during the DOPLIGHT '87 experiment is unusual for a field research program in terms of the a priori design rigor and extent of the forecast evaluation. Some of the results are surprising while others are more or less as expected. If one can generalize Weiss's (1977) view that verification of SELS severe weather forecasts tends to be better on days with major outbreaks of severe weather, then the DOPLIGHT '87 forecasts may represent a lower bound on the skill that one might expect in a typically active severe weather season.

Perhaps the most surprising result of the verification is the unexpected similarity between the dichotomous "go" forecasts of the advance outlook and those of the noon outlook. Although the noon outlook was, indeed, the best forecast, the margin of improvement over the advance outlook was rather modest. This suggests that large-scale evidence for strong convective activity typically is available well before the event, at least during the spring in Oklahoma. This has important implications for at least convectively oriented field programs in the future, since it indicates that important decisions about operating the project's data collection systems made the evening before an anticipated operational day have about as much success as waiting until the last possible moment, which is typical of many field research projects (Doswell et al. 1986).

It is of some interest to speculate on why the morning update was the least skillful of the three "go" forecasts. Figure 4 shows a time line for the forecast day during DOPLIGHT '87. We note that the morning sounding data arrive just before the 0900 L morning update is due. Further, the limited-area, fine-mesh model (LFM) forecast package (the earliest of the numerical model forecasts) arrives only *after* the morning update has been issued. This problem is exacerbated by the change to daylight savings time, since relative to local time, all these forecasting tools arrive an hour later during daylight time. Thus, the morning update typically was produced with little chance to digest the new data and no chance to see the most recent model output. Perhaps another factor of significance was the presence of convective "debris" from nocturnal storms, which often confuses the diagnosis of the situation but typically dissipates by late morning.

On the other hand, the best forecast was the noon outlook, which had the benefit of extra time to consider the latest data and model output. By that time of the day, moreover, the influence of any nocturnal convection normally has dissipated. If one is to expect any improvement over a forecast made the previous night, sufficient time must be allowed for the forecaster(s) to diagnose the data adequately (see Doswell et al. 1986).

If the DOPLIGHT '87 forecasts are any guide, it appears that further improvements in forecasting convective weather are likely to be rather modest without enhancing substantially the available database. Perhaps an important benefit to this verification exercise is that it establishes a rough baseline forecasting skill, against which we may be able to measure the impact of new observing and analysis technology. The relative lack of severe weather during the experiment, however, restricts generalization of these results.

The DOPLIGHT '87 forecasters generally had little or no experience with using polychotomous probability forecasts. It is now apparent (recall Fig. 2) that nonoperational forecasters should be given some basic training in how to use probability forecasts. Generally, our forecasters tended to underforecast the "go" events, often using intermediate probabilities when they should have been using 100% values. However, some of the 100% forecasts they *did* issue failed to verify. At least some of the explanation for these deficiencies is the

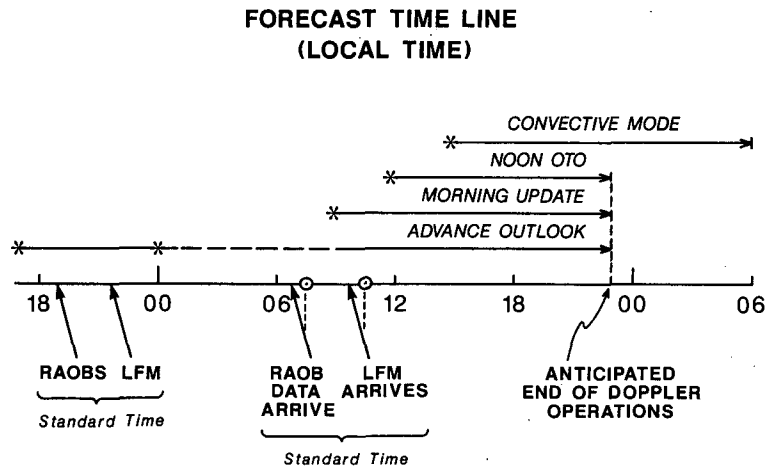

FIG. 4. Time line for the DOPLIGHT '87 forecasts. Times indicated along the line are given in local station time, while the arrows indicate when the sounding data ("raobs") and the first operational forecast model output ("LFM") normally arrive at the OKC office during central standard time (add 6 h to CST to obtain UTC). The dashed lines ending in circled dots show the change in arrival time when the change is made to central daylight time. Above the time line are indicated the issue times (asterisks) and valid times (solid lines) for the various forecast products.

diversity of experience within the forecast team, leading to different approaches to subjective probability estimation. Also, if forecasting is likely to have a significant role in determining the success or failure of a field project, considerable attention should be given to training the forecasters in all aspects of forecasting. Based on the DOPLIGHT '87 experience (two lost forecast days early in the season) we recommend that field project forecasting begin well before the actual operation of the project, in order to get the forecasters familiar with the process *before* the evaluation period begins.

It appears that the mesocyclone forecasts may exhibit some skill even with this group of diversely experienced forecasters. However, the limited occurrence of mesocyclones during the experiment dictates that this part of the forecasting effort be repeated in the future, in order to obtain more cases.

The convective mode forecasts turned out to be flawed in important ways, despite our efforts at careful a priori design. Our preliminary conclusion is that there is, perhaps, some modest skill in mode forecasting (a TSS of 0.36). Clearly, there is a need to know whether or not the convective mode can be forecast successfully, both in operations and in convective field programs. Future experimental forecasts of convective mode type need more careful design and forecasters should be trained carefully about the definitions of the modes offered.

Although this experiment is largely within the context of a research program, it was conducted in collaboration with an operational forecast office. Further, some of the forecast products were designed to explore possible new operational products that have been suggested for future operational implementation. Thus, the outcome of the experiment has implications for operational forecasting as well as field research programs.

In general, the convective forecasting procedures employed by the National Weather Service (i.e., the "convective outlook, watch, and warning") depend implicitly on the assumption that prediction of a convective event becomes more precise as the time of the event approaches. Our experiment did not include forecast products comparable to watches and/or warnings, which are more time- and space-specific products than outlooks. If watches and/or warnings are to verify well, the largest scale, outlook-type products certainly should improve as the time of the event approaches. Since we did not observe a steady improvement of the outlooks with time, this suggests that the underlying assumption may be unwarranted. Although more information about the developing situation is available as time goes by, this "ready-set-go" process suffers from two defects. First, the scale of the data does not change, while the important physical processes that determine the sensible weather tend to decrease in scale as the time of the event approaches (Doswell 1987). Second, the knowledge available to aid the forecaster in predicting those mesoscale (and smaller) processes is lacking, perhaps largely because the data on small scales historically has been absent.

Although DOPLIGHT '87 did not incorporate a very wide range of possible new operational technologies, it was clear that adding new information to the standard fare produced no dramatic increase in either forecast skill or real-time understanding of unfolding events. It appears that many of the new products remain inadequately understood in a forecasting mode. This tends

to produce confusion and uncertainty on the forecaster's part, because of lack of product credibility. In fact, with some products it was unclear exactly how one should interpret the output (e.g., derived products from clear-air operation of the Doppler radar, or quasi-geostrophic diagnostics). While we did not encounter "information overload", it seems increasingly evident that experiments of this type should be conducted prior to introducing new systems. Thus, we find ourselves in agreement with the conclusion of MacDonald et al. (1988) that forecasting will not improve automatically when new, enhanced data systems are implemented (although for somewhat different reasons). As noted in Doswell (1986), the introduction of new technology does not translate into instant improvement in forecasting skill.

Our forecasters did surprisingly well in comparison to those of SELS. We already have indicated some reasons why this may be so, and we wish to emphasize that our exploratory results do not resolve the speculation that local forecasters generally can do as well at severe weather outlooks and watches as severe weather specialists in a national center. Sanders (1986) found similar results when comparing MIT forecasts of temperature and precipitation to the forecasters at the Boston forecast office of the National Weather Service: the student/faculty forecast consensus was generally equal to or, in some cases, superior to the operational forecasts. Nevertheless, Sanders noted that the MIT forecasters had a limited set of forecast products for a single city, whereas the operational forecasters have a much more diverse set of forecast products, and are responsible for an entire state.

The decrease in quality of our morning update relative to the advance outlook, however, is primarily due to the lack of time to absorb the morning sounding information and to see the newest model output derived from it. If the forecast cycle (phased with the local time zone) is not well matched with the data cycles (phased with universal coordinated time), our results indicate that the product tends to suffer. Although there is a need for forecasts at fixed local times, if they can not be done with a reasonable amount of time to assimilate the lastest data, then it is unreasonable to expect those forecasts to verify well.

5. Concluding comments

Herein and in Flu89, we have introduced some verification displays and measures which may be somewhat unfamiliar to operational forecasters. The contingency table offers the most effective method for initial forecast verification. It gives a simple and easily understood picture of forecasting success and failure, which can serve as the starting point for examination of the reasons for success and failure. The extent to which the table is diagonally dominant is a crude measure of what is going right in the forecasts; as or more important are the off-diagonal cells which represent forecasts gone awry. This knowledge can point one in the proper directions for improving the predictions. What do the incorrect forecasts have in common with each other? How do they differ from the correct forecasts? Are the false alarms largely due to the quality of the observations? These questions and the subsequent evaluations should lead directly to improvements in the forecasts.

The contingency table is also a good basis for application of statistical analysis tools, including summary measures of skill, ROC diagrams, etc. It is tempting to use statistical analysis, calculate summary measures, and generate diagrams without ever looking at the meteorology behind the successes and failures. This does not resolve the primary issue posed by a verification exercise in the first place: the search for improvements in the forecasts. This task is much more difficult and time consuming than generating summary verification measures. Therefore, we have deferred such an effort with respect to the DOPLIGHT '87 forecasts to a later date.

Acknowledgments. We wish to extend our thanks to Mr. Mark Antolik of NSSL for his efforts in converting the ACs to dichotomous forecasts, and to Dr. Preston Leftwich of the Techniques Development Unit at NSSFC, for supplying us with verification data for the ACs. An important contribution to the forecasting part of the project was the cooperation and patience of the operational forecasters at the National Weather Service Forecast Office in Norman, led by Dr. Ken Crawford. The constructive criticisms of Dr. Fred Sanders and the anonymous reviewers are also appreciated. Finally, we are grateful for the dedication and enthusiastic support of all the participants in DOPLIGHT '87, without which this experiment could not have taken place.

APPENDIX A

Selected Traditional Verification Measures

As Hughes (1980) has noted, the "universally accepted verification score for probability forecasts is the score of Brier (1950)." While the events we are scoring are dichotomous (i.e., they either occur or they don't), the probabilities are not. Hence, the appropriate score is the half-Brier score, which employs the simple forecast difference

$$D_i = (F_i - O_i) \tag{A1}$$

for each individual forecast, where F_i is the forecast probability and O_i is the probability of the corresponding observed event (either zero or unity) for the ith case. The half-Brier score, B, is then the mean of the squared differences in (A1), or the mean square error of the forecasts, for a set of N cases,

$$B = \frac{1}{N} \sum_{i=1}^{N} D_i^2. \tag{A2}$$

The B score has a range of (0, 1), with zero corresponding to perfect forecasts, since it is actually a measure of the forecast *error*.

We also have used the so-called skill score (S), in the interest of providing a measure familiar to many operational forecasters. The S score is defined (see Sanders 1963) to be

$$S = \frac{100(B_c - B_f)}{B_c}, \qquad \text{(A3)}$$

where B_c is the B score for a constant climatological probability forecast and B_f is that derived from (A2). If one does not know the climatological probability for some event, it is possible to use the observed, or sample frequency to calculate the S score for any particular set of forecasts. This is what we have done for the DOPLIGHT '87 forecasts.

Finally, we have determined the overall bias (b) for the probability forecasts according to Hughes (1980)

$$b = \frac{100(R_f - R_o)}{R_o}, \qquad \text{(A4)}$$

where R_f is the average of the forecast probabilities and R_o is the observed frequency of events during the experiment. The bias score is designed to detect any systematic under- or overforecasting, with zero being ideal.

APPENDIX B

Evaluation of a $k \times k$ Contingency Table

The reader should consult Flu89 for definitions of the terms POD, FAR, CSI, POFD, and TSS. If one wishes to extend the contingency table evaluation methods to the $k \times k$ case ($k > 2$), let POD_i and FAR_i be the POD and FAR for the ith forecast category (e.g., the different convective modes). In a straightforward extension of the 2×2 table cell contents notation (as in Flu89),

$$\text{POD}_i = \frac{n_{ii}}{n_{\cdot i}}, \quad \text{FAR}_i = \frac{1}{n_{i\cdot}} \sum_{j \neq i}^{M} n_{ij}, \qquad \text{(B1)}$$

where M is the number of observed categories. Knowing the POD and FAR for any particular category, it is easy to determine the associated CSI (see Donaldson et al. 1975).

The TSS can be generalized to the $k \times k$ case as described in appendix A of Flu89. As in the 2×2 case, one must find the expected value for the cells in the table under the assumption that the cell contents are due to random chance, subject to the constraint that the totals remain unchanged along the table margins. In particular, for the TSS we are concerned with the table's trace, or the sum of its diagonal cells, after subtracting the expected value from each diagonal cell. For the DOPLIGHT '87 convective mode forecasts (Table 10), for example, the expected value for cell (1, 1) is given by $41 \times (35/89) = 16.12$ (see appendix A in Flu89). Repeating this for all the diagonal cells and subtracting the expected values from the diagonal cells in Table 10 gives a trace of 22.62.

If the forecasts were perfect, the contents of the diagonal cells would be equal to the observed totals in each category (e.g., 35, 31, 4, 8, and 11, reading from right to left in Table 10), with all other cells having zero values. Thus, the marginal totals for the forecasts also would equal the marginal totals for the observed events. If one uses these marginal values to recompute the expected values in the same manner as before, and subtracts those expected values from the diagonal cells in this hypothetical contingency table of perfect forecasts, the trace is 62.17. Thus, the TSS value is 22.62/62.17 = 0.36.

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

——, and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology*, T. F. Malone, Ed., Amer. Meteor. Soc., 841–848.

Brownlee, K. A., 1965: *Statistical Theory and Methodology in Science and Engineering* (2nd Ed.) John Wiley and Sons, pp. 211 ff.

Dixon, W. J., M. B. Brown, L. Engelman, J. W. Frane, M. A. Hill, R. I. Jennrich and J. D. Toporek, 1985: *BDMP Statistical Software 1985*. Univ. of Calif. Press, 733 pp.

Dobryshman, E. M., 1972: Review of forecast verification techniques. World Meteorological Organization, Tech. Note No. 120, 51 pp.

Donaldson, R. J., R. M. Dyer and M. J. Krauss, 1975: An objective evaluator of techniques for predicting severe weather events. *Preprints, 9th Conf. Severe Local Storms*, Norman, Oklahoma, Amer. Meteor. Soc., 321–326.

Doswell, C. A. III, 1986: The human element in weather forecasting. *Nat. Wea. Dig.*, **11**, 6–18.

——, 1987: The distinction between large-scale and mesoscale contribution to severe convection: A case study example. *Weather and Forecasting*, **2**, 3–16.

——, R. A. Maddox, and C. F. Chappell, 1986: Fundamental considerations in forecasting for field experiments. *Preprints, 11th Conf. Weather Forecasting and Analysis*, Kansas City, MO, Amer. Meteor. Soc., 353–358.

——, and D. W. Burgess, 1988: On some issues of United States tornado climatology. *Mon. Wea. Rev.*, **116**, 495–501.

Flueck, J. A., 1986: Principles and prescriptions for improved experimentation in precipitation augmentation. *Precipitation Enhancement—A Scientific Challenge*, R. R. Braham, Ed., Boston: Amer. Meteor. Soc., 155–171.

——, 1987: A study of some measures of forecast verification. *Preprints, 10th Conf. Probability and Statistics in Atmospheric Sciences*, Edmonton, Alberta, Amer. Meteor. Soc., 69–73.

——, 1989: A study of prediction-verification methodology from the contingency table viewpoint. *Wea. and Forecasting*, **3**, In press.

——, and D. W. Reynolds, 1986: A forecast experiment on the prediction of cloud conditions suitable for treatment in the Sierra Nevada. *Preprints, 10th Conf. Weather Modification*, Arlington, VA, Amer. Meteor. Soc., 13–18.

Forsyth, D. E., D. W. Burgess, L. E. Mooney, M. H. Jain, C. A. Doswell III, W. D. Rust and R. M. Rabin, 1988: DOPLIGHT 87 Program Summary. NOAA Tech. Memo. ERL NSSL-101, 194 pp.

Hughes, L. A., 1980: Probability forecasting—Reasons, procedures problems. NOAA Tech. Memo. NWS FCST 24 [NTIS Accession Number PB80-164353], 84 pp.

Kelly, D. L., J. T. Schaefer, R. P. McNulty, C. A. Doswell III and R. F. Abbey, Jr., 1978: An augmented tornado climatology. *Mon. Wea. Rev.*, **106**, 1172–1183.

——, J. T. Schaefer and C. A. Doswell III, 1985: Climatology of nontornadic thunderstorm events in the United States. *Mon. Wea. Rev.*, **113,** 1997–2014.

MacDonald, K. A., M. Danks and J. D. Abraham, 1988: A short-range forecasting experiment conducted during the Canadian Atlantic Storms Program. *Wea. and Forecasting,* **3,** 141–152.

Maddox, R. A., 1980: Mesoscale convective complexes. *Bull. Amer. Meteor. Soc.*, **61,** 1374–1387.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30,** 291–303.

Murphy, A. H., and H. Daan, 1985: Forecast evaluation. *Probability, Statistics and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

Ostby, F. P., E. W. Ferguson and P. W. Leftwich, Jr., 1988: A strange tornado season. *Weatherwise,* **41,** 32–40.

Panofsky, H. A., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology.* Penn. St. University, 224 pp.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2,** 191–201.

——, 1973: Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.*, **54,** 1171–1179.

——, 1986: Trends in skill of Boston forecasts made at MIT, 1966–1984. *Bull. Amer. Meteor. Soc.*, **67,** 170–176.

Thompson, J. C., and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.*, **83,** 249–254.

Weiss, S. J., 1977: Objective verification of the severe weather outlook at the National Severe Storms Forecast Center. *Preprints, 10th Conf. Severe Local Storms,* Omaha, Amer. Meteor. Soc., 395–402.

Wilson, F. W., Jr., and J. A. Flueck, 1986: A study of the methodology of low-altitude wind shear detection with special emphasis on the LLWSAS concept. FAA Tech. Memo. DOT/FAA/PM-86/4 [NTIS Accession Number AD-A164939/1/XAB], 101 pp.